DOCUMENT RESUME

ED 200 608 TM 810 030

AUTHOR Brinzer, Raymond J.
TITLE New Directions in Matching Familiar Figures Test Research Resulting From Scoring and Item Analyses.
INSTITUTION West Virginia State Dept. of Education, Charleston.
PUB DATE Feb 79
NOTE 36p.; Paper presented at the Eastern Educational Research Association Conference (Kiawah Island, SC, February, 1979).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Conceptual Tempo; Difficulty Level; *Item Analysis; *Research Methodology; *Scoring Formulas; Test Items; *Test Reliability; *Test Validity
IDENTIFIERS *Matching Familiar Figures Test (Kagan)

ABSTRACT

The problem engendered by the Matching Familiar Figures (MFF) Test is one of instrument integrity (II). II is delimited by validity, reliability, and utility of MFF as a measure of the reflective-impulsive construct. Validity, reliability and utility of construct assessment may be improved by utilizing: (1) a prototypic scoring model that will enable development of MFF norms; and (2) item analyses (performed on MFF test items) results which will reveal good test items, reveal defective test items, provide a graphic display of item performance, explain the origin of the current imbroglio about MFF test reliability and validity, and indicate steps necessary to enhance MFF instrument and research integrity. The impulsive-deliberative score (ID Score) is discussed as a potentially better scoring procedure than the double median split procedure for users of reflective-impulsive category information. Despite the limitations and inchoate nature of the research presented in this paper, it would appear that research directed along similar lines would be in the best interests of the scientific method. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200608

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

New Directions in Matching Familiar Figures Test Research Resulting From Scoring and Item Analyses

Raymond J. Brinzer, Ph.D.
West Virginia Department of Education
Charleston, West Virginia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. J. Brinzer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 030

Paper presented at the Eastern Educational Research Association Conference, Kiawah Island, South Carolina, February 1979.

Prior to introduction of the Matching Familiar Figures Test (MFF) by Kagan, Rosman, Day, Albert, and Phillips (1964), classification of subjects as reflective or impulsive styled was accomplished by a variety of subjective measures (Hartshorne, May & Maller, 1929; Sutton-Smith & Rosenberg, 1959). Introduction of the MFF provided a method for quickly and objectively determining a subject's cognitive style.

The MFF is beset by reliability, validity, and utility probelems. A survey of the literature reveals the nature of the problems, and a more intensive examination reveals their source.

Although the MFF has been in use for approximately 15 years, little measurement analysis has been noted in the literature. This situation endures despite the appearance of a considerable MFF research effort. For example, an ERIC search conducted in September 1977 revealed a total of 102 studies, listed under the conceptual tempo description. Of the 102 studies approximately two major studies were listed as dealing with instrument relibility as a major emphasis, and eight major studies with validity in the same light.

An intensive review of MFF research literature covering the Educational Research Index, Psychological Abstracts, and Dissertation Abstracts revealed a plethora of MFF research. However, not one study to date has been noted that deals with the basic integrity[1] of the MFF as determined by a classical measurement approach to the instrument's behavior.

The reliability and validity problems of the MFF are related, in part, to the scoring and classification system (error rate-response latency/double median split) used in operationalizing the reflective-impulsive classification construct. In turn, the scoring and classification system has impeded standardization of the instrument.

Salkind (1977) has recently moved to develop norms and a scoring model (Salkind & Wright, 1977) for the MFF instrument. His completion of a study designed to produce normative information for the MFF constitutes a significant step in MFF research. The step, however, constitutes little more than an academic exercise when the integrity of the MFF instrument is considered in terms of reliability, validity, and utility. This situation obtains because crucial significant steps in MFF instrument development and refinement in the classical measurement sense have been omitted.

A pervasive theme begins to emerge when one considers the MFF in the perspective of its research and development. That theme is characterized by a field or functional fixedness that is typified by movement in a consistent direction set by MFF instrument parameters and concepts established by Kagan, et al., in MFF instrument development and scoring. This theme leads to the current problem in MFF research.

---

[1] The Random House Dictionary of the English Language defines integrity as the state of being whole, entire, undiminished, sound, unimpaired, or perfect in condition.

Statement of the problem

The problem engendered by the MFF situation is a highly complex one. It is, specifically, one of instrument integrity. Instrument integrity is delimited by validity, reliability, and utility of the MFF instrument as a measure of the reflective-impulsive construct. No action has been taken to date in order to rectify the MFF problem situation. Analysis indicates that the solution to the problem lies in a maximally effective MFF instrument, scoring, and classification system. Little issue would appear related to the construct itself. For supporting evidence review Hartshorne, May, Maller (1929); Cattell (1937), Murray (1938); Solansky, Lippitt, and Redl (1950); Sutton-Smith, Rosenberg (1959); and Kagan (1964, 1965).

Objectives

Objectives for this research endeavor were set by the nature of the problem. Specifically, they were intended to provide a direction to problem solution resulting in the highest validity, reliability, and utility of construct assessment possible. Attainment of the foregoing may best be attained through the following:

1.0 Introduction of a prototypic scoring model that will enable development of Matching Familiar Figures Test (MFF) norms.
Additionally, the model should:
- 1.1 Increase MFF efficiency.
- 1.2 Provide improved MFF test reliability.
- 1.3 Provide improved MFF test validity.
- 1.4 Provide individual test administration and interpretation capability.
- 1.5 Solve some of the problems and allay many of the criticisms stemming from use of the current scoring model.

2.0 Presentation of item analyses results performed on the MFF test items.
Item analyses results should:

- 2.1 Reveal the good test items.
- 2.2 Reveal the defective test items.
- 2.3 Provide a graphic display of item performance.
- 2.4 Explain the origin of the current imbroglio about MFF test reliability and validity.
- 2.5 Indicate the steps that must be taken in order to enhance MFF instrument and research integrity.

Review of the Literature

As mentioned previously in this paper, a limited amount of research deals with the reliability and validity of the MFF Test. Of that research Hall and Russell (1974) researched the divergent and convergent validity of conceptual tempo. The researchers found that no divergent validity existed for conceptual tempo on the MFF, Word Recognition Test, Raven Coloured Progressive Matrices, and Peabody Picture Vocabulary Test. This would tend to indicate that the trait is generalizable across tasks, as consistent response time tendency emerged on the researched tasks. Reliability for errors (number correct) on the MFF was reported lowest of the four instruments used in the study, with mean improvement being less than one item (the MFF was lowest here also). The authors reported that the low reliability questioned the double median split classification procedure.

Block, Block, and Harrington (1974) reported on the MFF Test as a measure of reflection-impulsivity. The authors reported that Kagan defines the concept in narrow terms, but applies it in a broad general sense. Additionally, they indicated the "the evidence for the construct validity of the MFF was sparse, often inconsistent, and sometimes irrelevant" (p. 612). The authors indicated that their intent was to:

> describe the discrepancy between Kagan's conceptualization ... and his operationalization of reflection-impulsivity; ... assess the construct validity of the MFF ... and ... present a representative portion of ... data ... that bears on ... the MFF situation (p. 612).

Ault, Mitchell, and Hartmann (1976) reported that Kagan's original reliability assessment was listed at .62 for latency, while error score reliabilities were cited in the .23 - .43 range. Although the authors stated that the low reliabilities could be due to a cognitive tempo stability lack, it would appear that the item performance of the MFF would account for a considerable degree of instability. Readers interested in reliability (test-retest/internal consistency) are referred to the article for an intensive discussion. The researchers closed with the statement that (MFF's) "validity has been demonstrated over a wide variety of tasks which measure cognitive development" (p. 230). The researchers recommend larger sample sizes, appropriate research designs, and statistical treatments as methods capable of making work with the present form of the test possible.

Egeland and Weinberg (1976) investigated the psychometric credibility of the MFF. They reported reclassification differences favoring reflective subjects with an 80/90 percent reclassification rate and a 50/56 percent rate with impulsives for one second-grade study. Other reclassification information was provided, and readers are referred to the article for a more comprehensive treatment. Additionally, the researchers cited Block, et al's., characterization of impulsives as fearful, inhibited, as consistent with their own interpretation. The authors stated that "the findings raise issue with the typical practice of labeling subjects solely on the basis of MFF Test data" (p. 489).

The authors recommended use of a linear time-error composite rather than the typical nonlinear approach inorder to avoid inherent double median split misclassification problems. In closing the researchers wrote:

> While one might question the premature acceptance of the MFF as a psychometric procedure for operationalization of the reflection-impulsivity construct, one might also urge caution and restraint in prematurely rejecting the test as an operational measure of reflection-impulsivity because its psychometric underpinnings have been uninvestigated (p. 490).

Salkind (1977) introduced normative tables at the 1978 AERA Convention. The normative information included descriptive data, means, and standard deviations for errors and response latencies by age; correlations of errors and latency by age and sex; and percentile rank information. The norming population encompassed the 5-12 year-old age range. Salkind's undertaking constitutes a crucial step in the MFF development as a measure of reflection-impulsivity, however the step preceded an array of more fundamental steps necessary to increase instrument integrity prior to normalization. Salkind's undertaking was a significant step beyond the functional fixedness pattern of much of the existing MFF research, and should ultimately engender a significant contribution to instrument integrity.

Scoring The MFF

The Double Median Split

MFF test results have been consistently scored by the double median split procedure devised by Kagan. This procedure typically involves adminstering the MFF test to a group, or groups, of subjects, then ranking all response latencies from lowest to highest, and all error rates from lowest to highest. The median (P_{50}) for the response latencies is then calculated. Likewise, the median error rate is obtained. Then, each subject's test results are examined to determine classification as reflective or impulsive. Typically, 35 percent of a group is reported classified as reflective, 35 percent impulsive, while the remaining 30 percent is unclassified (i.e., fast accurate and slow inaccurate; Hall and Russell, 1974, p. 933). Fast accurates and slow inaccurates are those subjects who fall above the median on response latency and error rate, or below the median on both measures respectively.

A variety of characteristics may be attributed to a double median split scoring procedure. Some of the characteristics appear positive, while others appear in a more negative light. Only the more salient negative characteristics will be discussed here. Typically, they involve the following:

1. Measures are group dependent, i.e., relative to specific groups. Technically, a group of reflective subjects could, by virtue of individual processing differences, be artifically classified as reflective or impulsive. A case in point would be one in which several classes are independently classified as reflective or impulsive. All reflective subjects from the several classes could then be combined and the double median split procedure applied. The reflectives could then be classified as reflective or impulsive. Additionally, unclassifieds could ostensibly achieve new classificatory status via the double median split procedure. This classification variance would appear to have serious implications for the double median split scoring procedure. On the other hand, a standard scoring procedure using $\bar{X}$'s, Ranges, SD's, and an index score combining response latency and error rate via a ratio, would appear to preclude many problems attributable to the present scoring system.

2. The double median split procedure assumes that a specific within group distribution exists. This implicitly negates the possibility of the construct being normally distributed within the population, and atypically distributed within specific groups.

3. Sex, age, SES, and other performance differences have been reported in the research literature. The double median split scoring procedure would not appear to demonstrate the potential capacity to systematically treat these differences, as they would tend to be offset by the groups themselves. Development of specific normative data for these groups, on the other hand, would appear to place them in a more appropriate classificatory perspective. For instance, performance differences by sex, race, or SES might appear more salient, and valid, from a classificatory perspective when these factors are controlled. Additionally, differences between or among groups, e.g., by sex, might be more validly attributed to specific group characteristics or performance.

4. Measures may vary considerably. The double median split scoring procedure may result in a considerable variation in classification as a result of seemingly inconsequential score differences. For example, the classification percentages for the research data used as the basis for this paper demonstrate the following differences for the 6-8 grade levels (see Appendices A, B, and C for specific data):

Table 1

Double Median Split Classification Variance for a Limited n at Three Grade Levels[a]

			Classification					
	Grade	n[b]	Reflective		Impulsive		Unclassified	
normal upper limit of the MFF Test	6	15	7	46.6%	7	46.6%	1	6.68%
	7	21	9	42.8%	10	47.2%	2	9.5%
	8	33	9	27.2%	13	39.9%	11	33.3%

[a] see Appendices A, B, and C for particulars.

[b] n differences may constitute a defect in this calculation (e.g., the extremely small n for Grade 6); nevertheless, the objective here is to demonstrate a potential defect. This defect would possibly be amplified due to n variance.

The reflective variance demonstrated across the 6-8 grade levels ranges from 27.2-46.6 percent. The impulsive variance 39.9-47.2 percent, while unclassifieds range from 6.68-33.33 percent. The implications of this variance (19.4, 7.3, and 26.6 percent respectively), attributable to the double median split scoring procedure, ought to be fairly obvious. It would appear that this aspect of scoring constitutes a significant portion of the MFF reliability/validity imbroglio.

5. Time expenditures for scoring are considerable in the case of the double median split. Time economy will be discussed later (see p. 9).

6. Due to individual differences, score variability, and specific characteristics of the double median split group-relative system -- individual administration of the MFF is not possible. It appears that scores must always be related to the specific groups/s.

7. An analysis of the state of the art concerning the MFF would appear to strongly indicate that an attempt to standardize the present instrument is subject to the limitations discussed in this paper and elsewhere. However, this is not an attempt to discredit such an undertaking, as the implications of moving in this direction are in themselves momentous. Additionally, Thorndike and Hagen (1977, p. 94) have stated that "a test with relatively low reliability will permit us to make useful studies of and draw accurate conclusions about groups", which appears to be the case concerning the MFF.

The ID Score

Response latency and error rate are essential components of the reflective-impulsive construct. Historically, response latency and error rate have been treated via the double median split scoring procedure mentioned previously. The group dependence, potential variability, and inadequacy for standardization of the double median split procedure indicate that another scoring procedure would better serve researchers, psychologists, school counselors, and other potential users of reflective-impulsive category information. An ideal scoring procedure would have the potential for individual administration, standardization, increased reliability and validity, time economy, and concomitant trait identification and analysis. This scoring procedure would appear to combine response latency and error rate into an index that could then be related to a classical measurement framework, including such aspects as X's P_{50}'s, SD's, SEM's and item analyses (power, discrimination, reliabilities, and validities). In this manner a more effective MFF instrument could be developed--resulting in far greater reliability, validity, and instrument utility (see Appendix G for recommended selected norming controls). Such a procedure and results are made possible through the ID Score (impulsive-deliberative; named after H. A. Murray, an early researcher in the area).

The ID Score is the ratio of $\overline{X}$ response latency to $\overline{X}$ error rate. It is obtained by the algebraic formula:

Table 2

ID Score Formula

ID Score =	$\frac{\frac{RL}{n}}{\frac{ER}{n}}$	= $\frac{TRL}{TER}$[a] , or	$\frac{\text{Sum Response Latencies}}{\text{number of subjects}}$	=	Total[a] Response Latency	
			$\frac{\text{Sum Error Rates}}{\text{number of subjects}}$		Total Error Rate	

[a]preferred computation due to convenience/rapidity

The formula produces a score that would appear to be a somewhat better measure of individual impulsivity-reflectivity. This is due to the group interactive nature of the double median split procedure. Directions for calculation of an ID Score, and a facsimile ID Score sheet/directions, are located in Appendices D, E, and F respectively. Examination of these appendices should give the reader a somewhat better idea of the potential for scoring ease and standardization that is characteristic of the ID Score. A calculation using actual data is entered on the score sheet for review (see Appendix F).

The ID Score excluding the aspect of a zero (0) base, linear trend, and open upper end, would appear to have considerable potential for normalization

[2]Early research results indicate that impulsive ID Scores generally range below ten.

and resolution of many of the problems presently attributable to the double median split scoring system. A considerable amount of research will have to be conducted with this procedure in order to build the research groundwork necessary to soundly establish the appropriate standardization process and illuminate the potential pitfalls.

Increased MFF Efficiency

Test efficiency may be increased in a variety of ways. Additionally, efficiency ought to be approached from different perspectives, such as examiner, examinee, instructions, procedures, scoring, and results interpretation and use.[3] Some, or all, of the preceding aspects would seem to have an effect upon test efficiency, and in combination that effect might tend to be dramatic.

Efficiency is viewed here from a scoring perspective. Consequently, data recording, computation, and time economy are primary considerations on the one hand. On the other hand -- item function would appear to be involved in test efficiency -- however -- it appears more properly relegated to the reliability and validity realms. Subsequently, efficiency is synonomous here with utility.

The smallest group (grade 6, n = 15) was selected in order to get an idea of MFF scoring efficiency. Following are the efficiency particulars from a scoring perspective for one versed male scorer:

Table 3

Scoring Efficiency Information for the Double Median Split and ID Score Systems Respectively

System	n	Tabulation Time	P_{50} Time	Classification Time	Total Time	Efficiency Index
Double Median Split	15	Pretabulated (includes posting RL/ER's and determing individual Total	2' 44"	3' 17"	6' 61" (361")	2.01 times as long as the ID Score method (i.e., twice the time)
ID Score	15	Pretabulated (see above)	2'	59"[a]	2" 59" (179")	.495 or 50% of double median split scoring time (i.e., ½ the time)

[a] substituted a table similar in format due to the lack of developed tables.

[3] Virtually no conformity was noted in the scoring area other than use of the double median split.

On the basis of this limited trial conducted by the researcher, it appears that the ID Score system is considerably more economical in terms of time expended in the scoring process. Additionally, it is hypothesized that, as the n increases, the time advantage in favor of the ID Score will become even more pronounced -- due to the physical limitations of the double median split procedure. It may take fully twice as much, or more, time to classify subjects using the double median split routine. An extensive research investigation of time expenditures in both systems would appear to contain more definitive answers to any questions raised here.

Test Reliability and Validity

A test can have extremely high reliability and little or no validity. However, a test cannot be qualitatively valid with low reliability. Reliability is considered a necessary quality to validity.

Problems concerning the reliability and validity of the MFF text have been discussed earlier in this paper (Hall & Russell, 1974; Block, et al., 1974; Kagan, 1965; Egeland & Weinberg, 1976), as well as elsewhere, e.g., (Kagan & Messer, 1975).

Although reliability is necessary to validity, relatively low realiability, such as in the case of the MFF, does not disqualify a psychological construct. Relatively low reliabilities, however, dictate the nature of related research and justifiable interpretation of results (this concept was briefly covered in the section "The Double Median Split," p. 6). For instance, Thorndike and Hagen wrote that (partially quoted earlier)

> a test with relatively low reliability will permit us to make useful studies of and draw accurate conclusions about groups, but relatively high reliability is required if we are to have precise information about individuals (p. 94).

The current MFF scoring system would appear to have little value for individual difference research. However, increased reliability -- and consequently, validity -- should improve the quality of research findings and generalizability regarding groups, as well as enable justifiable research and educational decision-making in the area of individual differences.

Consideration of Kagan's (1965) original reliability assessment data in the perspective of individual and group reliabilities indicated that the .62 response latency finding reported and the .23 - .43 error score range would have the following implications for percent of reversals with repeated testing (i.e., retesting):

Table 4
Approximate Reversal[a] Percentage Chance Figures[b] for MFF Retesting Using Selected Reliability Scores[c]

Category	Single Individuals	Groups (X of 25)	Group (X of 100)
Response Latency	@.62R = 1/3 or 32.5%	1/83.3 or 1.2%	less than 1/2500 or less than .04%
Error Score	@.23R = 1/2.19 or 45.15%[d]	1/3.28 or 30.45%	1/4.05 or 24.55%[d]
	@.43R = 2/5 or 40.3%	1/9.17 or 10.9%	1/142.85 or .7%

[a]reversal, i.e., chance for being classified reflective, impulsive, or unclassified one time and changing classification upon retesting.
[b]Thorndike and Hagen source (1977, p. 93).
[c]Kagan reliabilities source (see p. 4).
[d]approximations based on interpolation of the Thorndike - Hagan table. This procedure may be invalid. The intent is to communicate general implications, not exact data.

It is believed that -- due to the synergistic nature of reliability and validity -- and the synergistic nature of the response latency and error rate in the measurement of reflectivity-impulsivity in the double median split -- the cooperative action of the low reliabilities for response latency and error rate is such that the total deficiency in terms of reliability and validity is maximally less then the low reliability of the response latency taken separately, and minimally more than the error rate reliabilities taken separately, i.e., would fall somewhere in the middle area (circa .475, the midpoint, may be too high an estimate due to the peculiar relationship of these reliabilities). Consequently, an idea of the nature of the reliability- validity controversy in the MFF area can be gained, and the necessity of approaching the problem in a classical measurement fashion, and the latter's implications for improving test reliability and validity, appreciated.

Item Analyses Results

As stated earlier, no item analysis results dealing explicitly with the MFF instrument have been noted in the research literature. The MFF test was individually administered by the researcher in this particular instance to a select remedial reading population drawn (n = 227, or 10%) according to the following criteria out of a total N of approximately 2200: Remediation classification was based upon a qualifying score on the Metropolitan Reading Readiness Test in Grade 1 of low C, D, or E and teacher/supervisor judgement; and a cummulative deficit of three months per grade level (e.g., 1.7 grade 2, 2.4/3, 3.1/4, 3.8/5, 4.5/6, 5.2/7, and 5.9/8) in the comprehension section of the Gates-MacGinitie Reading Test over the 2-8 grade levels and teacher/supervisor judgement. Basic MFF test results follow in Table 5, shown below.

Table 5

Response Latency and Error Rate Ranges for an Experiment with an n of 227

Grade	n	Response Latency R	Error Rate R	
1	37	1.9 - 70.5	4 - 29	
2	34	2.48 - 60.64	4 - 28	
3	37	3.17 - (41.75[a])	(4)- 24	
4	28	2.22 - 49.38	3 - 23	
5	23	3.16 - 62.33	3 - 19	normal upper limit of the MFF Test (elementary edition) used in this study
6	15	5.0 - 40.77	1 - 14	
7	21	3.85 - 73.5	2 - 17	
8	33	3.88 - 55.38	3 - 13	

[a]Extreme score cut (369.5 with 1 error)

Note. Scheduling and other factors influenced the middle school sample used for this study (Grades 6-8). Actually, the information is somewhat different than it appears due to the fact that approximately one-half (two of four participating elementary schools) of approximately 340 remedial reading elementary subjects representing all of the elementary remedial reading subjects were tested, while all of the 6-8 grade students were tested. The 6-8 grade student sample, however, did not constitute the entire middle school remedial reading population.

Definite trends become apparent upon examination of the data in Table 5. Those trends consist mainly of a general increase in minimum response latency over the grade levels 1-8 (1.9-3.88), an initial decrease in maximum response latency over grade levels 1-3 (70.5-41.75), then an apparent increase at the intermediate level (49.38-62.33). Upper elementary or middle school ranges are somewhat erratic. Once again, it is believed that the low n (15/20 at Grades 6/7) may contribute to this display.

An examination of the error rate range indicates that, generally speaking, the minimum number of errors decreases over 1-8 grade level range (4-3), while the maximum number of errors decreases also (29-13). The trends demonstrated in this data would appear to be in line with expectancy. An increased n in all cases would appear to be a necessary factor in future experiments along similar lines. Additionally, the clear trend for decreasing response latencies with age would appear to be somewhat contradicted by this particular set of tabular data. It would appear that cognitive maturation, in part might contribute to a decreasing response latency with age, especially in the case of reflective subjects. The increasing response latency - age relationship is readily apparent only in the case of low response latency subjects for this particular data.

Good and defective test items. An item analysis performed on individual items across the 1-8 grade levels reveals a variety of item performances. In viewing item difficulty levels, e.g., a classical measurement approach would indicate that items functioning systematically could be expected to demonstrate performance in the .30 - .70 range. Furthermore, the slope of this performance ought to be positively linear if the items are functioning effectively and their performance is related to cognitive maturity.

Startling results emerged from the initial item analysis performed on the test data. These results indicate that approximately eight of the items (#' 1, 4, 5, 6, 7, 8, 9, and 12) are defective in this particular instance, while four of the items (2, 3, 10, and 11) may be termed good items. Defective items are defined as those items that:

1. Are too easy -- such as item #5 which demonstrates a range of 61 - 91 percent over grade levels 1-8.

2. Are too hard -- such as item #12 which demonstrates a range of 31 - 27 percent over grade levels 1-8.

3. Demonstrate an unsystematic or sporadic slope, such as item #4 which traverses a range from 33 percent (Grade 1) to 62 percent (Grade 3) to 33 percent (Grade 8).

On the other hand, good items are defined as those items that demonstrate systematic slope across the grade levels, such as the slopes demonstrated by items 2, 3, 10, and 11.

A tabular display of item analysis results (difficulty level) that is supportive of the preceding item classifications is included in Table 6. Comments about the general nature of the items are included in the comments column.

Table 6

MFF Item Difficulty Data

General Comment	n = Grade =	36 1	36 2	37 3	29 4	22 5	13 6	21 7	33 8
Too hard	Item #1	.305	.222	.351	.241	.363	.692	.285	.393
Good	2	.388	.500	.594	.620	.500	.538	.666	.727
Good	3	.277	.636	.405	.448	.454	.461	.619	.575
Poor	4	.333	.333	.621	.344	.409	.538	.428	.333
Too easy	5	.611	.638	.729	.724	.909	.769	.857	.909
Too easy	6	.222	.611	.540	.620	.636	.769	.857	.848
Too hard	7	.388	.416	.324	.310	.318	.846	.333	.545
Too hard	8	.138	.277	.243	.172	.363	.384	.523	.545
Erratic	9	.250	.250	.459	.310	.590	.307	.523	.272
Good	10	.250	.305	.351	.517	.545	.461	.666	.666
Good	11	.305	.500	.513	.310	.636	.846	.761	.696
Too hard	12	.305	.166	.135	.241	.181	.307	.285	.272

normal upper limit of the MFF Test used in this study

Item performance graphic display. Perhaps the most dramatic description of specific item performance for this data can be obtained by graphically displaying each item. For this purpose an ideal, or artificial, item curve has been included. Although developmental trends do not always follow the ideal, or linear, it is anticipated that a rather systematic slope ought to be the case in terms of effective item functioning and developmental differences. This is indeed the case as concerns the good items. However, the performance of the defective items appears rather self-explanatory. Ideal, good, defective, composite, and comparative (good - defective - ideal) item curve tables are included according to the following schedule:

1. Table 7 - ideal item curve for eight grade levels (p. 15).
2. Table 8 - ideal item curve for six grade levels (p. 16).
3. Table 9 - defective item curves (p. 17).
4. Table 10 - composite defective item curve (p. 18).
5. Table 11 - good item curves (p. 19).

6. Table 12 - composite good item curve (p. 20).

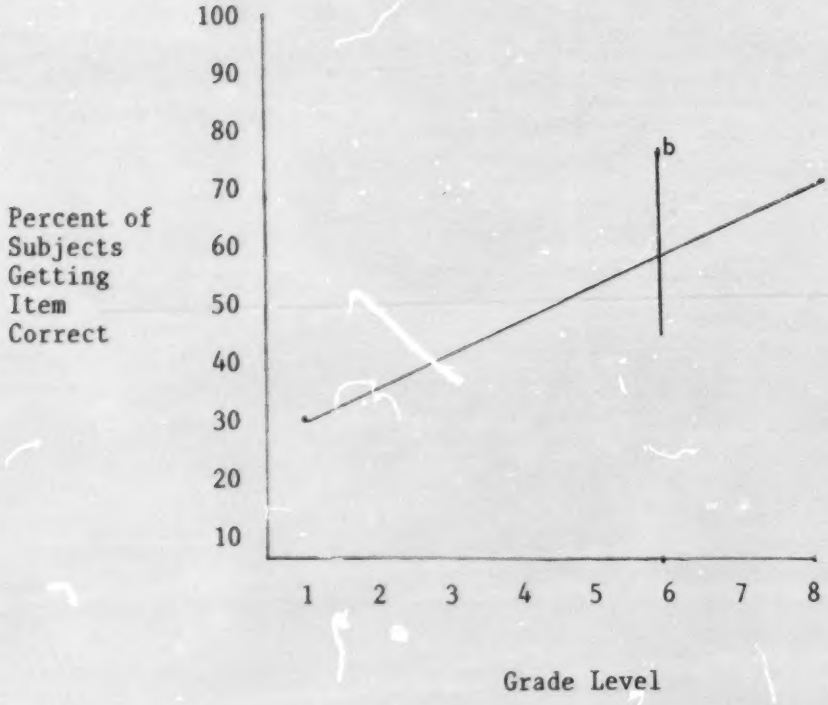
7. Table 13 - good - defective item curves for comparison/contrast (p. 21).

8. Table 14 - percent differences for good - defective - ideal items (p. 22).

The tables follow (see pp. 15-22).

Table 7

Eight Grade Level
Ideal Item Curve[a]

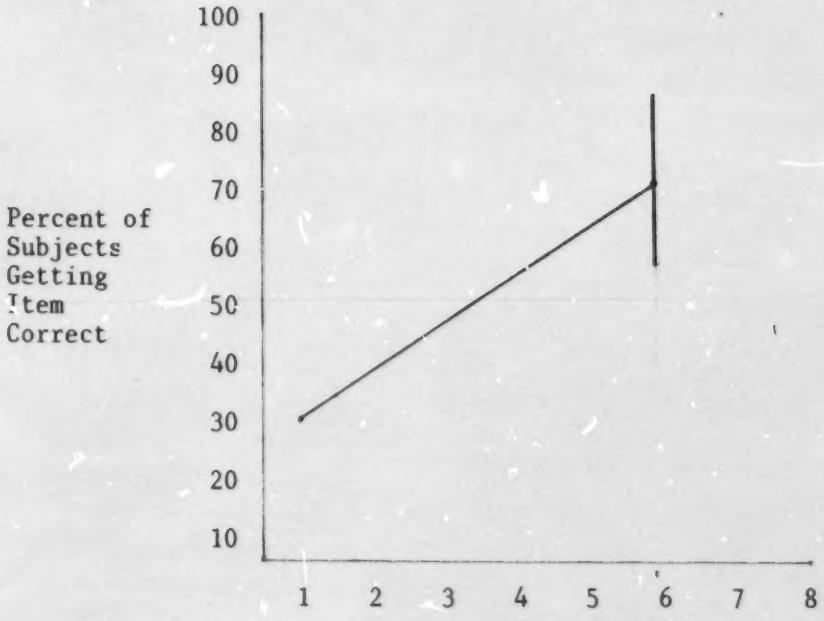


Range = 30 - 70 percent
Rate of change = 5.714% per grade level[a]
Grade level/percent = 1/30, 2/36, 3/41, 4/47, 5/53, 6/59, 7/64, 8/70

[a]computed independent of developmental surges
[b]normal upper limit

Table 8

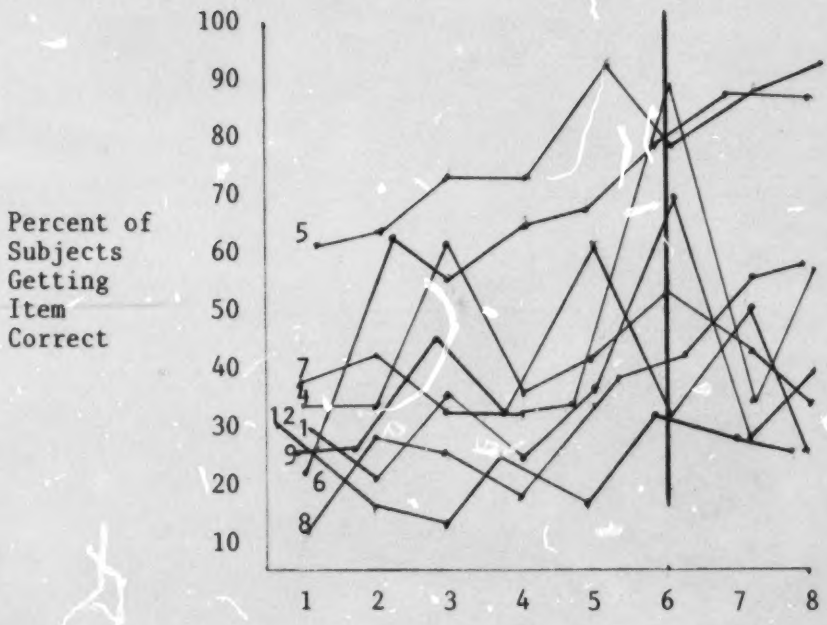
Six Grade Level
Ideal Item Curve



Range = 30 - 70 percent
Rate of change = 8 percent
Grade level/percent = 1/30, 2/38, 3/46, 4/54, 5/62, 6/70

Table 9

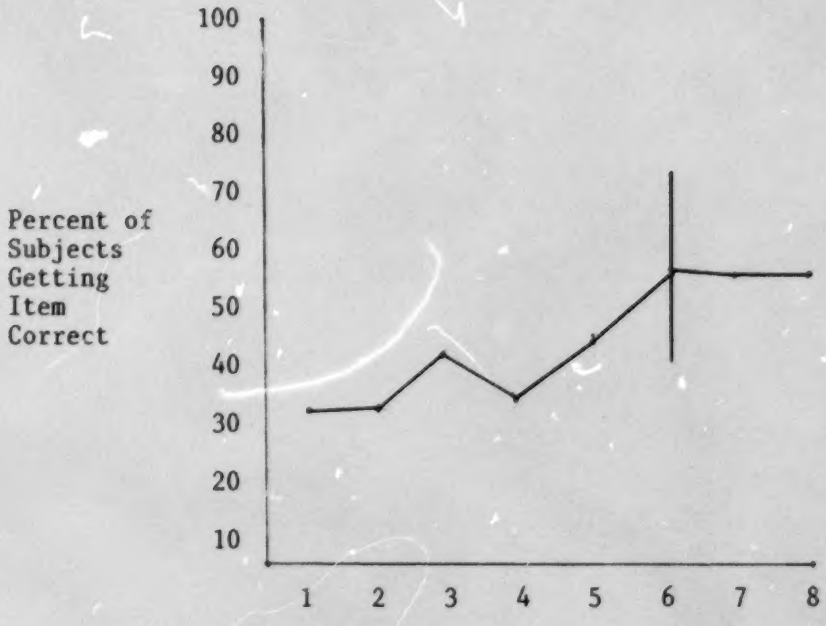
Defective Item Curves



Range = 33 - 51 percent
Defective items = 1, 4, 5, 6, 7, 8, 9, 12
Grade level ranges/percent = 1-6/33-58, 1-4/33-37, 5-8/47-51, 7-8/51-51

Table 10

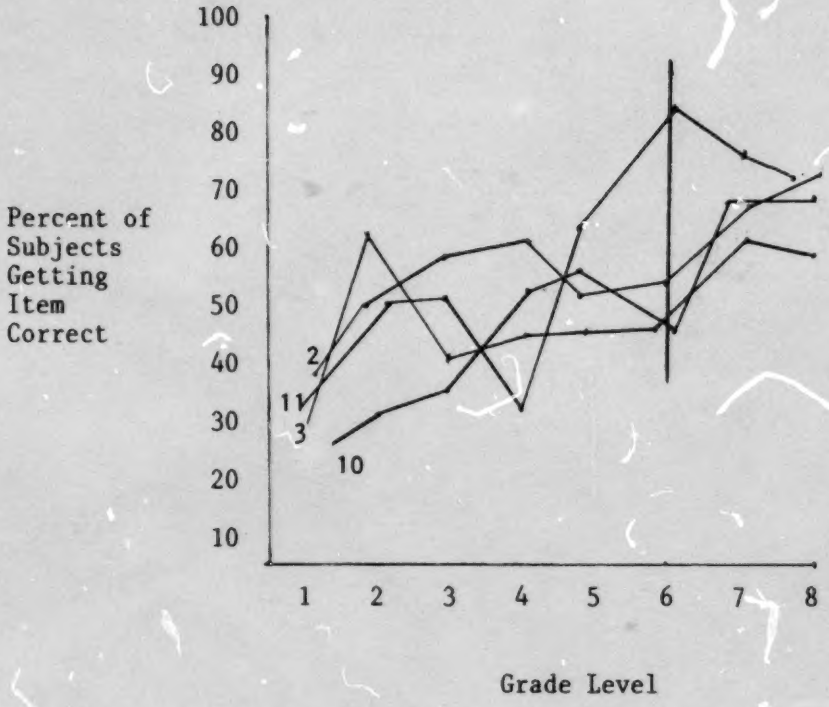
Composite Defective
Item Curve



Range = 33 - 51 percent
Defective items = 1, 4, 5, 6, 7, 8, 9, 12
Grade level ranges/percent = 1/33, 2/36, 3/43, 4/37, 5/47, 6/58, 7/51, 8/51

Table 11

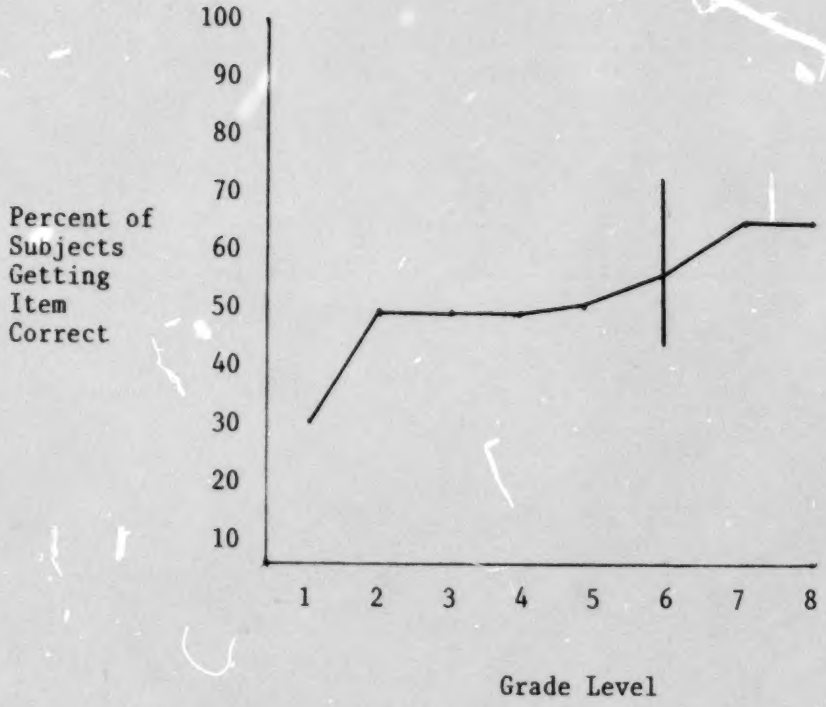
Good Item Curves



Range = 31 - 67 percent
Good items = 2, 3, 10, 11
Grade level ranges/percent = 1-6/31-58, 1-4/31-47, 5-8/53-67, 7-8/68-67

Table 12

Composite Good
Item Curve



Range = 31 - 67 percent
Range/Good items = 2, 3, 10, 11
Grade level/percent = 1/31, 2/49, 3/47, 4/47, 5/53, 6/58, 7/68, 8/67

Table 13

Good - Defective Item
Curves/Curve Differences

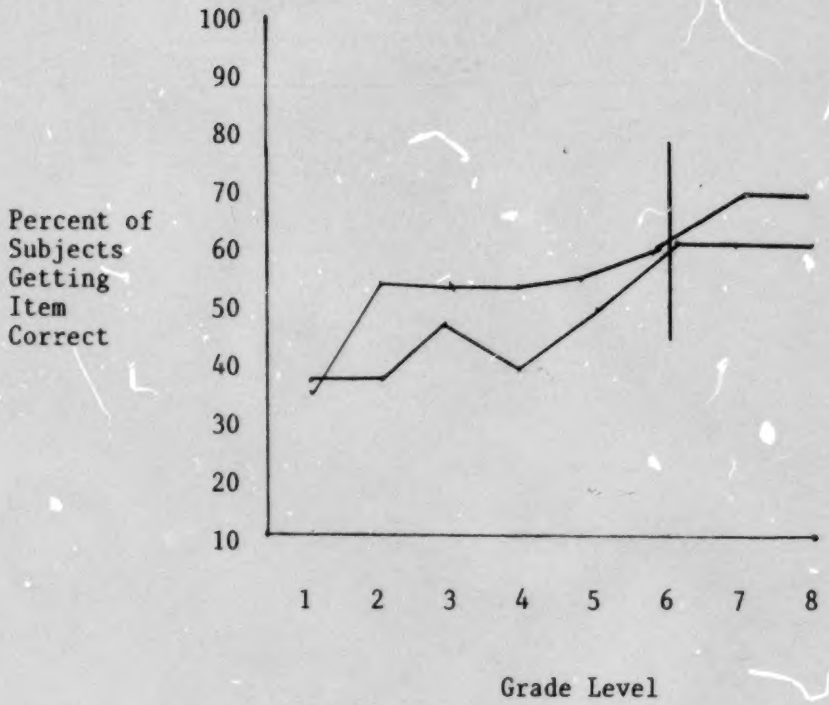


Table 14

Percent Differences for
Good - Defective and Ideal -
Good - Defective Items

Grade Level	1	2	3	4	5	6	7	8	Total		
Good Items	31	49	47	47	53	58	68	67			
Defective Items	33	36	43	37	47	58	51	51			
Difference	-2	13	4	10	6	0	-17	-16	= -2[a]		
Ideal Slope	30	36	41	47	53	59	64	70			
Good Difference	1	13	6	0	0	-1	4	-3	= +20[b]	=	64[d]
Defective Difference	3	0	2	-10	-6	-1	-13	-19	= -44[c]		

[a]Total Difference Good - Defective = -2
[b]Total Difference Good - Ideal = +20
[c]Total Difference Defective - Ideal = -44
[d]Total Difference Ideal - Good/Ideal - Defective = 64

Origin of the Current MFF Reliability - Validity Imbroglio

It would appear that the origin of the current MFF imbroglio is attributable to a variety of causes. One of the most basic causes may reside in the lack of historical knowledge about the construct itself. Indeed, many researchers attribute the concept of reflective-impulsive response style to Kagan, but it is the more objective operationalization of the concept via the MFF and its scoring system that is attributable to Kagan (along with a formidable body of construct research and conceptual development). Any discussions of validity and reliability, or the construct, must necessarily consider early research and development completed on the topic shortly after the turn of the century, and intermittently down to the present time. This lack of historical knowledge foments a potential flaw for much of the contemporary thought, research, and criticism of the reflective-impulsive construct.

Another basic cause of the current reflective-impulsive imbroglio would appear to reside in the double median split scoring procedure discussed earlier. Much intensive research on the implications of the double median split scoring procedure would appear to be in order.

The MFF Test items themselves would appear to be latent sources of text reliability and validity problems. The potential authenticity of this statement increases when the item graphic displays presented earlier in this paper are considered, and the potential implications for validity and reliability are considered along the lines of reliability implications for reversals of score classifications reported by Thorndike and Hagen (see p.10). The implications of the item analysis performed here would appear to have profound implications for the present form of the MFF Text itself, as well as research conducted using this instrument, and the future direction of MFF research.

Steps Necessary to Enhance MFF Text/Construct Versatility, Research, and Integrity

A variety of steps would appear necessary when one considers the scope and nature of the problems besetting the MFF area. Initially, it appears -- based upon the research and development of Hartshorne, May, and Maller; 1929, Murray; 1938, Kagan, et al., 1964, that the construct itself is sound. Any claims concerning or questioning the construct must assume the burden of disclamatory proof -- which appears no mean task in face of the research evidence. Consequently, many of the steps that must be taken have been suggested or stated in this paper, tacitly or explicitly. They are a logical consequence of problems raised or issues broached. A reiteration will be advanced at this point so that a frame or general perspective may be advanced. The components of that frame include a need for:

1. A thorough and comprehensive analysis of the historical development of the reflective-impulsive construct. This analysis must not start in the late 1950's, or with Kagan, but should trace the development of the construct as far as is possible.

2. A thorough and scientific item analysis of the MFF Test items. This analysis should include item difficulty, item discrimination, and test validity.

3. Development of a new, or revision of the old, MFF Test using classical measurement principles.

4. Development of a new scoring system. This system should provide the MFF (revised or new test) with the potential for increased reliability, validity, and utility (e.g., individual administration). Additionally, this system should be economical from a time standpoint. Such a system appears in Appendices D, E, and F of this paper; Appendix G includes a matrix and comments listing selected norming considerations.

5. Development of norms for the revised, or new, MFF Text. These norms should likewise support individual administration (#4 preceding), and provide more substantive data about specific group and individual characteristics. Specific reliability and validity figures should be included. The norms should be developed classically in a SD format, with the SEM concept included. This may ultimately result in, 68.26 percent of a group being classified as unclassified (and subsequently possibly more amenable to experimental treatment), while the remaining 31.74 percent would fall in progressively more reflective or impulsive categories. This approach is almost a direct reversal of the double median split scoring procedure. Note the following comparisons:

Table 14

Student Percent Distributions
for Two Scoring Systems

Scoring System	Percent Reflective	Percent Impulsive	Percent Unclassified
Double Median Split	35	35	30
ID Score Format	15.86	15.86	68.26
Scoring System Difference	19.14	19.14	-38.26

Additionally, Appendix G includes a matrix and comments listing selected norming minimums.

6. Comparison/contrast of the double median split - ID Score procedures/ systems and their implications for research and practice. This includes the practice of neglecting response latencies past the initial error.

7. A review of more salient early studies in the perspective of subsequently established normative tables.

8. An analysis of MFF scoring trends in order to determine if increasing reflectivity or impulsivity is a characteristic of contemporary trends.

9. Comparison/correlation of reflective-impulsive groups with specific traits/characteristics in the area of cognitive style and achievement. The new scoring system would increase controls and the opportunity for such studies.

10. Calculation of reliabilities and validities for the double median split scoring procedures as well as the ID Score procedure. A comparative analysis of the implications of both systems ought to add substantive knowledge to the reflective-impulsive construct research area.

11. Investigation of the double median split low error score reliability problem (M = .23/F = .24; Messer, 1970) in the perspective of ID Score reliability. The ratio dimension of the ID Score may have positive implications for score stability in the face of its response latency relationship.

12. The growing movement towards normalization of a test that is fraught with reliability and validity problems ought to be examined. For example, Salkind (1977) has developed norms based on existing research data obtained during investigations by other researchers. This researcher had considered such a move, but initial item analysis results obtained during preliminary test data analyses were considered sufficient to preclude such action. Subpopulation score differences, coupled with low reliabilities, further compound this problem.

13. Development of an annotated bibliography for the reflective-impulsive area. Such a bibliography should include validity, reliability, reading, sex, and other differential factor study citations.

Limitations

A considerable number of limitations exist for this paper. Some of the more salient ones are:

1. The MFF Test is used interchangeably with the reflective-impulsive construct.

2. The depth of analysis conducted here has been somewhat superficial. Computation of item discrimination, reliability, and validity indices, along with other statistics, would seem to add much valuable information on which to base judgements.

3. The comprehensive focus of this paper is a limitation closely related to #2.

4. The low n would appear to be a severe limitation.

5. Use of a remedial reading sample constitutes a severe limitation. However, if two-third's of the MFF Test does not function systematically for a population of this nature -- it would appear that generalizability to similiar groups would be highly suspect. This aspect combines with #'s 2 and 4 to preclude such an undertaking with this data.

6. The selection process for Grades 6-8 constitutes a deficiency and may account for some of the erratic data at these levels.

7. The zero point - open end of the ID Score comprises a problem that needs investigated. The resulting distributions will be positively skewed. The nature of the response latency - error rate relationship, however, indicates that this is not a serious deficiency -- if a deficiency at all.

8. Inclusion of Grades 7 - 8 for a test that has a normal upper limit of 12 years of age is enlightening from one perspective, but a bona fide limitation from another.

Despite the limitations and inchoate nature of this research and its findings -- it would appear that research directed along the lines suggested in this paper, and by other researchers, would be in the best interests of the scientific method. This is especially true when the current state of the art is viewed in the perspective of a classical measurement approach to sound test development.

Appendix A

Random ordinal scores (unrelated)

Grade 6
n = 15 Related scores

Response Latency	Error Rate		Response Latency		Error Rate	Classification
46	1		46	-	23	Impulsive
46.5	3		73	-	11	Impulsive
52	4		104	-	5	Reflective
60	4		110	-	10	Unclassified
73	4		194.5	-	1	Reflective
92	5		112	-	3	Reflective
100	7		92	-	9	Impulsive
104	8	Median (P_{50})	163	-	4	Reflective
110	9		46.5	-	8	Impulsive
112	10		157	-	7	Reflective
116	10		100	-	14	Impulsive
126.5	11		116	-	4	Reflective
157	12		60	-	12	Impulsive
163	14		126.5	-	4	Reflective
194.5	23		52	-	10	Impulsive

P_{50} = 104/8 (Actual)

Impulsives = 7 = 46.6%
Reflectives = 7 = 46.6%
Unclassifieds = 1 = 6.68%

Total = 15 = 100%

Note: Impulsive = at or below P_{50} on RL - at or above on ER
Reflective = at or above P_{50} on RL - at aor below on ER
Unclassified = above or below P_{50} on RL and ER

The Unclassified subject (110/10) would be called a slow inaccurate. The line separating such subjects would appear to be exceedingly find indeed, insofar as this particular case is concerned.

Appendix B

Grade 7
n = 21

Response Latency	Error Rate		Response Latency		Error Rate	Classification
62	2		93	-	13	Impulsive
65.5	2		109.5	-	6	Reflective
66	4		65.5	-	17	Impulsive
83	6		98	-	10	Impulsive
84	7		154	-	8	Reflective
85.5	7		83	-	10	Impulsive
89.5	8		66	-	8	Impulsive
93	8		103	-	8	Reflective
95	8		142	-	8	Reflective
98	8		89.5	-	11	Impulsive
99	8	Median (P_{50})	102	-	2	Reflective
102	8		133.5	-	10	Unclassified
103	9		156.5	-	4	Reflective
109.5	10		243	-	7	Reflective
133.5	10		147	-	2	Reflective
134	10		99	-	10	Impulsive
142	10		95	-	8	Unclassified
147	11		134	-	7	Reflective
154	11		84	-	11	Impulsive
156.5	13		85.5	-	9	Impulsive
243	17		62	-	8	Impulsive

P_{50} = 99/8.25 (actual)

Impulsives = 10 = 47.2%
Reflectives = 9 = 42.8%
Unclassifieds = 2 = 9.5%

Total = 21 = 100%

Note: The Unclassified subjects (95/8, 113.5/10) demonstrate what appear to be marginal differences to earn the label fast accurate and slow inaccurate respectively.

Appendix C

Grade 8
n = 33

Response Latency	Error Rate		Response Latency		Error Rate	Classification
50.5	3		221.5	-	4	Reflective
53	3		95	-	7	Impulsive
54	4		96	-	3	Reflective
57	4		83	-	9	Impulsive
62	5		113	-	7	Reflective
66.5	5		184.5	-	5	Reflective
66.5	5		62	-	10	Impulsive
67	5		113	-	10	Unclassified
72.5	5		76	-	6	Unclassified
76	5		77	-	12	Impulsive
77	6		57	-	10	Impulsive
81	6		54	-	14	Impulsive
82	6		81	-	7	Impulsive
83	7		72.5	-	6	Unclassified
95	7		66.5	-	7	Impulsive
95.5	7		101.5	-	9	Unclassified
96	7	Median (P_{50})	149.5	-	7	Reflective
100	7		189	-	5	Reflective
100	7		119.5	-	9	Unclassified
101.5	8		180.5	-	5	Reflective
113	9		82	-	6	Unclassified
113	9		66.5	-	5	Unclassified
119.5	9		100	-	9	Unclassified
130	9		67	-	13	Impulsive
134.5	9		158	-	3	Reflective
149.5	10		134.5	-	7	Reflective
155	10		95.5	-	4	Unclassified
158	10		130	-	5	Reflective
178.5	11		50.5	-	13	Impulsive
180.5	12		53	-	9	Impulsive
184.5	13		155	-	11	Unclassified
189	13		178.5	-	8	Unclasified
221.5	14		100	-	5	Reflective

P_{50} = 96/7.083 (actual)

Impulsives = 13 = 39.39%
Reflectives = 9 = 27.2%
Unclassifieds = 11 = 33.33%

Total = 33 = 100%

Note: Eyeball analysis of the Unclassified in Grade 8 reveals a pattern of marginal fast accurates and slow inaccurates, while several subjects demonstrate more pronounced differences (e.g., 155/11 would appear to be a bona fide slow inaccurate, while 101.5/9 would not).

Appendix D

ID Score Computation

Directions: Use these steps with Appendix F in order to obtain the ID Score.

Step

1. ∈RL column 7.
2. ∈ER column 8.
3. Divide ∈RL (column 7) by ∈ER (column 8) to obtain the ID Score (column 11).

Appendix E

ID Score Sheet Directions

Directions: To use this sheet with Appendix F complete all steps in order. Step numbers equate to Appendix F notations.

Step

1. Enter subject's name of record.
2. Enter subject's age.
3. Enter subject's sex.
4. Enter subject's grade.
5. Enter subject's test date.
6. Enter subjective examiner observation, i.e., classification as reflective or impulsive obtained during testing.
7. Enter response latency in seconds to first response.
8. Enter error rate as errors occur.
9. Enter error order as errors occur.
10. Enter relevant notes as incidents occur during the test situation.
11. Compute the ID Score by dividing the column 6 total (response latency) by the column 7 total (error rate). See Appendix D.
12. Obtain the subject's birth date from the records and enter it.
13. Compute chronological ages as of the test date.
14. Enter the number of years in school (this information should be obtained from the record and may not agree with grade placement due to retention).

Note: Information noted on the ID Score sheet is considered minimal. The following specifics are added for explanatory purposes:

Step

6 - Subjective observations (SO) may someday be compiled and correlated in order to provide important information about individual scores and the norming system. SO, however, requires considerable experience with test administration to gain the proficiency that would seem necessary to function effectively in this area.

8 - Although the total number of errors per item is a consideration -- only the initial response latency is used in ID Score formula calculation. General examinee response latency behavior after the initial error

should accordingly be entered in the Notes column (#10). This total errors - initial response latency situation exists for the double median split scoring procedure also, and would seem to warrant investigation.

9 - Error order information may be used at a latter date to provide valuable item discrimination information in the classical measurement vein.

10 - Notes (column 10) should include explanatory and test relevant behavioral observation data.

11 - The ID Score is used to classify subjects as reflective or impulsive. This score will have little formal value until the test has been replaced or revised, and normed accordingly.

12/13 - Birth data and chronological age information should be obtained from the subject's record. This information may be used to provide chronological age control for later score analysis.

14 - Years in school information is intended to cover retentions. Later analysis of students with additional years in school (retentions) may provide valuable test performance/behavioral performance data for this group.

Appendix F

ID Score'Sheet

[1]Name _____ [4]Grade _____

[2]Age _____ [14]Years in School ____

[3]Age _____ [6]SO _____ [5]Date _____

[12]Birth Date _____

[13]Chronological Age _____

Item	Response Latency[7]	Error Rate[8]	Error Order[9]	ID Score[11]	Notes[10]
Sample A		0			Noise coming from cafeteria
Sample B		0			
1. House	13.5	3	3, 6, 5		
2. Scissors	10	0			Said pointed - didn't see
3. Phone	18.0	0			
4. Bear	8.0	5	5, 3, 6, 4, 2		
5. Tree	20.0	1	6		
6. Leaf	4.0	2	6		
7. Cat	12.5	0	5, 1		
8. Dress	18.0	4	2, 3, 1, 4		
9. Firaffe	15.0	2	6, 2		
10. Lamp	10.5	2	1, 6		
11. Boat	15.5	0			
12. Cowboy	11.0	3	2, 3, 1		Random directional
Total	156.0	22			Unclassified x double-median
ID Score				7.09	

Appendix G

Suggested ID Norming Matrix

Note: Acquisition of the following information should provide effective test norms. The test to be normed, however, must be sound prior to undertaking the norming process.

Grade / Sex / Subject n	September/February[a]															
	1		2		3		4		5		6		7		8	
	M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
1. $\overline{X}$ age (total)																
2. $\overline{X}$ age (retentions)																
3. $\overline{X}$ age (regulars)																
4. $\overline{X}$ IQ (total																
5. $\overline{X}$ IQ (retentions)																
6. $\overline{X}$ IQ (regulars)																
7. Response Time $\overline{X}$ R (total)																
8. Response Time $\overline{X}$ R (retentions)																
9. Response Time $\overline{X}$ R (regulars)																
10. Error Rate $\overline{X}$ R (total)																
11. Error Rate $\overline{X}$ R (retentions)																
12. Error Rate $\overline{X}$ R (regulars)																
13. ID Score Information for 1 - 6 preceding ($\overline{X}$'s, R's, SD's, SEM's)																

x suburban subjects
x rural subjects[b]
x inner-city subjects
x race
x cross-national groups
x retentions

Additional information would be useful. This information might include reading test scores (vocabulary/comprehension/total) for the specific groups obtained by a concurrent administration, e.g.

[a,b] Norms for a September and February administration would greatly improve the utility of the data as well as the data interpretability process. Reliabilities would automatically follow from such a norming procedure. It appears that a grade level n of approximately 400 subjects (200 boys/200 girls) would be necessary to soundly undertake the norming process for a specific population segment, such as rural subjects, with a September/February administration.

References

Ault, R.L., Mitchell, C., and Hartmann, D.P. Some methodological problems in reflection-impulsivity research. Child Development. 1976, 47, 227-231.

Block, J., Block, H.H., and Harrington, D. Some misgivings about the Matching Familiar Figures Test as a measure of reflection-impulsivity. Developmental Psychology. Vol. 10, No. 5, 611-632.

Cattell, R.B. Measurement versus intuition in applied psychology. Character and Personality. 1937, 6, 114-131

Egeland, B., Weinberg, R. The Matching Familiar Figures Test: A look at its psychometric credibility. Child Development. 1976, 47, 483-491

Hall, V.C., and Russell, W.J.C. Multitrait-multimethod analysis of conceptual tempo. Journal of Educational Psychology. 1974, Vol. 68, No. 6, 932-939.

Hartshorne, H., May, M.A., and Maller, J.B. Studies in service and self-control. New York: The MacMillan Company. 1929.

Kagan, J., Rosman, B.L., Day, D., Albert, J., and Phillips, W. Information processing in the child: Significance of analytic and reflective attitudes. Psychological Monographs. 1964, 78 (1, Whole No. 578).

Kagan, J. Impulsive and reflective children: Significance of conceptual tempo. In J.D. Krumboltz (Ed.), Learning and the Educational Process. Chicago: Rand McNally, 1965, 133-161.

Messer, S. Reflection-Impulsivity: Stability and school failure. Journal of Educational Psychology, 61, December 1970, 487-490.

Murray, H.A. Explorations in personality. Oxford University Press, 1938.

Polansky, N., Lippitt, R., and Redl, F. An investigation of behavioral contagion in groups. Human Relations: Studies towards the integration of the social sciences. Vol. III, No. 4, 319-348.

Salkind, N.J. The development of norms for the Matching Familiar Figures Test. University of Kansas, paper. Lawrence, Kansas, 1977.

Salkind, N.J., and Wright, J.C. The development of reflection-impulsivity and cognitive efficiency: An integrated model. Human Development. 1977, 20, 377-387.

Sutton-Smith, B., and Rosenberg, B.G. A scale to identify impulsive behavior in children. The Journal of Genetic Psychology, 1959, Vol. 95, 211-216.

Thorndike, R. and Hagen, E. Measurement and evaluation in psychology and education, 4th edition. John Wiley & Sons, 1977, 93-94